Jailbreak Attacks and Defenses Against Large Language Models: A Survey

Year: 2024 | *Citations: 189* | *Authors: Sibo Yi, Yule Liu, Zhen Sun*

Abstract

Large Language Models (LLMs) have performed exceptionally in various text-generative tasks, including question answering, translation, code completion, etc. However, the over-assistance of LLMs has raised the challenge of"jailbreaking", which induces the model to generate malicious responses against the usage policy and society by designing adversarial prompts. With the emergence of jailbreak attack methods exploiting different vulnerabilities in LLMs, the corresponding safety alignment measures are also evolving. In this paper, we propose a comprehensive and detailed taxonomy of jailbreak attack and defense methods. For instance, the attack methods are divided into black-box and white-box attacks based on the transparency of the target model. Meanwhile, we classify defense methods into prompt-level and model-level defenses. Additionally, we further subdivide these attack and defense methods into distinct sub-classes and present a coherent diagram illustrating their relationships. We also conduct an investigation into the current evaluation methods and compare them from different perspectives. Our findings aim to inspire future research and practical implementations in safeguarding LLMs against adversarial attacks. Above all, although jailbreak remains a significant concern within the community, we believe that our work enhances the understanding of this domain and provides a foundation for developing more secure LLMs.